DOCUMENT RESUME

ED 271 482 TM 860 303

AUTHOR Busch, John Christian; Jaeger, Richard M.
TITLE Judges' Background, Attitudes and Information as Concomitants of Recommended NTE Test Standards.
PUB DATE Apr 86
NOTE 41p.; Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement; Adults; Cognitive Dissonance; Educational Attitudes; Educational Background; Individual Characteristics; *Judges; Minimum Competency Testing; Occupational Tests; Questionnaires; *Standardized Tests; *Standards; *Teacher Evaluation; *Test Validity
IDENTIFIERS *National Teacher Examinations; *North Carolina; Standard Setting

ABSTRACT

This study examined the contribution of five categories of variables to the standard-setting recommendations of a representative sample of 241 judges recommending standards for the General Knowledge and Communication Skills Tests of the National Teacher Examinations (NTE) for the state of North Carolina. It was based on the assumption that questions which addressed the contributions of concomitant variables may have important methodological and theoretical implications for understanding standard-setting practices. Five categories of concomitant variables included: (1) judges' background; (2) judges' attitudes; (3) judges' confidence in their recommended standards; (4) judges' understanding of information and procedures of standard-setting; and (5) the extent to which judges perceived they utilized directions and information. Standards were computed by summing individual item-probability judgments. Analysis of responses to an administered questionnaire indicated that participants generally agreed with the need to validate and set standards for the NTE. A majority of participants felt confident about their judgments, and they made use of information intended to help them in making recommendations, when such information was available. Results were generally consistent with predictions based on theory: judges who support neither standard setting objectives nor the specific use of the NTE might recommend more lenient standards. It was concluded that relationships between most characteristics of judges and their recommended test standards were modest. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED271482

JUDGES' BACKGROUND, ATTITUDES AND INFORMATION AS CONCOMITANTS OF RECOMMENDED NTE TEST STANDARDS

John Christian Busch and Richard M. Jaeger
University of North Carolina at Greensboro

TM 860 303

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J.C. Busch

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco CA, April 1986

INTRODUCTION

This study examined the relative contribution of five categories of variables to the standard-setting recommendations of a representative sample of 241 North Carolina judges. These judges recommended standards for the National Teacher Examinations (NTE). The study was based on the assumption that questions which addressed the contribution of concomitant variables may have important methodological and theoretical implications for understanding standard-setting processes. The five categories of concomitant variables included: judges' **background** variables (such as gender, and profession), judges' **attitudes** (e.g., toward the need to set standards in teacher education, and toward the use of standardized tests to set standards, etc.), judges' **confidence** in their recommended standards, the extent to which judges **understood** the information and procedures of standard setting, and the extent to which judges perceived that they **utilized** the directions and information provided in the standard-setting process.

This paper describes part of the results of an 18 month study designed to determine the validity of, and appropriate standards for, the General Knowledge and Communication Skills Tests of the NTE for the State of North Carolina. A detailed report on the final procedures used to investigate test validity and establish recommended standards for the test was provided to the North Carolina State Department of Public Instruction in 1984 (Jaeger & Busch, 1984a). In that study, judgments were obtained from a statewide sample of 306 public school teachers and college and university faculty members. Teachers who were graduates of

approved North Carolina teacher education programs were sampled, using accepted scientific procedures, from school systems in each of the three major regions of the state. Faculty were sampled from all North Carolina higher education institutions with approved teacher education programs. Each judge spent an entire day taking a subtest of the NTE, carefully reviewing each subtest and following a rigorous day-long judgment procedure through which s(he) provided recommendations for an appropriate passing score. Judges used an iterative, modified Angoff procedure in which they were given the opportunity to discuss their recommendations. Another section of this report provides a detailed description of the sampled judges and the procedures they followed.

During recent years standard setting methodology has been the focus of much research attention. There is no consensus among the community of scholars about a single preferred methodology. We are still at the point of gathering fundamental information.

A review of the literature did not identify any studies in which the concomitant variables identified above, or similar variables, were measured nor has the relationship of these variables to recommended standards been explored.

Many researchers have suggested that standards are best established by using representatives of different populations on the standard setting panels. For example Popham and Yalow (1984) state that, in the case of teacher competency tests, "the most likely judges are teachers, citizens and higher education representatives" (p. 8). Many reports describe the characteristics of judges but do not address the issue of relationships between the

characteristcs of judges and the standards they recommended.

One of the essential features of the standard setting procedure used in this study was the iterative nature of the process; the judges were invited to reconsider their initial recommendations following a structured group discussion. This procedure was described in an earlier report (Jaeger & Busch, 1984b). The rationale underlying the Delphi procedure was described by Brandstatter et al.,(1982, p. 4) as:

> an attempt to recognize and combat the social sources of inefficiency in group problem solving...The idea is to separate persons and bring them together again in an orderly fashion, to secure the advantages long recognized for isolation...at the same time to use discussion for catching errors and exercising solution criteria which may not be evident to individuals alone.

Fitzpatrick (1984) has completed a valuable critical analysis of the social psychology literature which addresses the effect of group discussions and exposure to the group's opinions on panel members' judgments in standard setting. She indicated that group discussions often produce a shift of judges' opinions to a more extreme position; the shift is called "group induced polarization" (Myers & Lamm, 1976). Fitzpatrick identifies two explanations for "polarization." One interpretation states that group discussion provides norms for the group member who shifts his judgment to accommodate to those norms; a second interpretation focuses on the informational character of the process and states that the discussion provides an information base of arguments for a preferred position and, as a result, the judge learns and shifts position. She concludes that "when group members are, on the average, initially disposed to privately favor one side of an issue or the other, discussing the issue will polarize their

opinions" (p. 11). She indicates that when a group member simply learns the opinions of other group members, some polarization occurs, though it is less than the polarization which follows discussion.

She concludes that,

> Discussion among group members was noted .. .as a procedure that is thought to elicit informational influences through the exchange of arguments as well as social comparison processes through the inferences that group members make about others' positions in light of their arguments. Hence, some kind of discussion appears desirable to incorporate in the standard-setting process" (p. 16).

She indicates that two ways normative effects can be minimized are through private recording of members' opinions or by carefully guided and balanced discussion. Further she states that "... judges will be less likely to engage in social comparision when they have been given data or other objective information about the test performance they are being asked to judge" (p. 17).

Although there is not an extensive literature addressing the issue of the relationship of concomitant variables to standard setting, there is a rich and venerable literature in social psychology on group decision making and conformity. The literature includes the classic studies initiated by Asch (1951) on group conformity, that of Sherif (1947) on the autokinetic effect, and more recently the study of shift to risk (Stoner, 1961).

There is some support (Swap, 1984) for investigations which examine subject characteristics in group decision making : "The personalities, attitudes, and skills that members bring to the

group, and how their individual characteristics mesh with the needs of the group...are all important in determining the nature of group interaction and the quality of group decisions" (p. 46). The literature suggests that groups whose members possess higher status tend to promote greater conformity (Stephan & Stephan, 1985) although Brandstatter et al. (1982) suggest that it is difficult to separate group member expertise and confidence levels. The panels in the present study were composed of members with differential status. It was therefore possible to compare the standard-setting behavior of public school teachers with university faculty in order to address this issue. While Brown (1965) speculates that there would be some influence of gender on responses to Stoner-type problems, more recent commentators suggest "that sex and personality variables are weak predictors of conformity" (Stephan & Stephan, 1985, p. 155).

The study examined the following questions:

a. Descriptive Questions:

What were the **attitudes** of judges toward the screening of teacher education candidates? What were the **background** characteristics of judges? How **confident** were judges about their recommendations? How well did judges **understand** and **utilize** the information provided to them? What were the **standards** recommended by the judges?

b. Relational Questions:

Did judges who differed in **background**, **attitudes**, **confidence**, **understanding** and **utilization** of information recommend standards that differed in central tendency or

variability? What were the relationships of these variables if any, to the patterns of recommendations provided by the judges during the course of a multi-stage standard-setting process? Changes in recommended standards following judges' discussion of their standards and the provision of information about the consequences of standard setting were examined, as were relationships between those changes and concomitant variables. Also examined was whether such attitudes varied across NTE subtest for which standards were set.

Although this study is correlational and strict causal inferences are precluded, the study was intended to support the generation of hypotheses on factors that contribute to judges' standard-setting decisions. As work continues to be conducted on features of standard-setting processes, it is important to assess the relative contribution of subject-related characteristics as they interact with variables that may be within the control of psychometricians e.g., the standard-setting methods used.

METHODS

Characteristics and Selection of Judges:

The design used for sampling judges is described in detail in Jaeger and Busch (1982). It required that three types of judges be selected. Selection was to be completed by using a probability sampling procedure that ensured the unbiased representation of (a) statewide populations of public school teachers who were relatively recent graduates of approved teacher

education programs and (b) college and university faculty who contribute to the education of beginning teachers and who were employed in either discipline based departments or education departments or schools.

The design proposed that data be collected at three sites that were convenient to colleges, universities and school systems in the mountain, piedmont and coastal regions of North Carolina. Judges of all three types participated at the data collection site that was closest to their institution.

The data collection sites were selected so as to secure an effectively representative statewide sample of public school teachers who satisfied the criteria described above, without attempting the impractical task of sampling teachers from every school system in North Carolina. The one county that best represented each of the three regions of North Carolina on six demographic variables was selected from each region. Teachers were sampled from all school systems that were within a 35 mile radius of the selected county. Using this criterion, 48 school systems were identified as sources of public school participants. The State Department of Public Instruction provided sampling frames for all teachers who met the following five criteria: full time teacher; graduate of one of 44 approved teacher education programs in North Carolina; currently employed in a sampled school system; a holder of a valid teaching certificate as of 1983; initially certified no more than five years and no less than two years prior to 1983. Teachers were sampled using a stratified sampling procedure with region as the stratification variable, and proportional representation by region.

A detailed description of the development of the sampling frames for college and university faculty in discipline departments who taught undergraduate general education courses in subjects assessed by the General Knowledge and Communication Skills Tests of the NTE, and faculty in departments, schools and colleges of education, are described in detail in Busch and Jaeger (1983). The sampling design was intended to meet a number of objectives: First, to ensure representativeness, a probability sampling procedure that incorporated random selection of participants at the final stage of sampling was to be used. Second, institutions in each region were represented in proportion to the numbers of graduates from their teacher education programs in recent years. Third, institutions that traditionally served minority students were sampled at a higher rate than were institutions that traditionally served non-minority students. The number of judges of each type who were to serve on the review panels assembled for the seven multiple-choice sub-tests and the writing essay test of the NTE were determined in advance. College and university faculty reviewed subtests which were appropriate to the area of their teaching responsibilities. (For example, discipline based faculty from biology, physics, and chemistry departments and teacher education faculty who taught science education courses were selected to review the NTE science subtest). The sampling process is described in detail in Jaeger and Busch, (1984a). In total, 100 public school teachers and 206 college and university faculty served on the eight panels.

The procedures used to set standards on the essay writing

subtest and the other multiple choice tests differed; as a result this paper reports results only for the seven multiple choice NTE subtests. A total of 241 judges provided recommendations on appropriate standards for the seven subtests. The number of public school teachers and college/university faculty is shown in Table 1.

Insert Table 1 about here

Data Collection Procedures:

Each panel was instructed by a knowledgeable group leader who was supported by an assistant. These group leaders had professional training and responsibilities in the subtest area (e.g. Reading, Mathematics etc.) reviewed by their respective panels and were themselves trained by the principal investigators during a five- hour training session. All procedures were standardized and were conducted using detailed scripts, handout materials, and transparencies. The standard-setting activities were alternated with exercises used to assess the validity of the NTE subtests; data collection lasted a full day. Only the standard-setting and related activities will be described in the present paper.

The session began with a review of background information on the North Carolina Quality Assurance Program and the National Teacher Examinations. Judges then "took" the subtest which they were going to review. The standard-setting activity consisted of four major parts: Instruction in standard setting, an initial

standard-setting session, additional training, including a mini-test exercise, and finally, a second standard-setting session. Detailed description of the process can be found in the scripts contained in Jaeger and Busch (1984a). A modified Angoff standard-setting procedure was used. Each judge was asked to consider a minimally competent person applying for admission to their teacher education program; they then reviewed individual test items and estimated the probability that the hypothetical individual would answer each item correctly. Judges selected and marked their probability values on a specially designed mark-sense form selecting from a ten-choice range of probabilities. The machine-readable mark-sense form was then used with a Scantron Model 1200 form reader and a micro-computer to compute the initial recommended standard, on-site, for each judge. The standard was the sum of the item probabilities marked by the judge.

Following a break, the panelists were provided with a mini-test exercise which was used to instruct them on the interpretation of two types of "item probability" statistics. The first was the proportion of examinees in the total population of North Carolina teacher education applicants who responded correctly to each item. The second probability was the proportion of individuals getting the item correct, who, on the total test, answered between 30% and 50% of the items correctly. This second group was used to reference a "minimally competent" group. In addition, the panelists were instructed on the interpretation of a cumulative percentage curve, which allowed judges to determine the percentage of examinees statewide who would fail the test if the test standard were set at the judges' recommended

standards

During the second standard-setting session, the standard-setting forms from the first session were returned to each panelist, along with their recommended passing scores (based on their earlier probability judgments). The judges were also given the two types of item probabilities described above, in addition to the cumulative percentage distribution of scores on the subtest they were reviewing. These data were based on the spring 1983 statewide administration of NTE. Judges were then asked to reconsider their earlier item judgments, considering the information provided, and to complete another set of item judgments. The procedure was different from that used in the first standard setting session; judges were encouraged to discuss their ratings with other panel members in a structured leader-directed discussion before re-rating the items. To facilitate the discussions, the leaders were provided with information which summarized the first session ratings for each item. An attempt was made to obtain opinions from judges who assigned relatively "high" and "low" probabilities to the items during the first session. Discussions were focused on pre-selected subsets of items from each instrument which had been grouped on a logical basis (e.g., items which used the same "stem" information). Judges then repeated the standard-setting process, conceptualizing a minimally competent individual, and estimating the probability that the hypothetical individual would answer the item correctly.

At the conclusion of the day's activities, each judge completed a participant questionnaire. The questionnaire consisted

of twenty-five positive and negative statements regarding the judges' attitudes toward standard-setting and toward screening applicants to teacher education programs, the judges' confidence in their recommendations, judges' understanding of the information provided to them, the procedures used, and whether judges used and were influenced by the information provided. These items are reproduced in Table 2.

Insert Table 2 about here

Statistical Analysis:

Descriptive statistics were used to summarize the judges' recommended standards for each of the subtests; the results of the judges' responses to each of the questionnaire items are summarized in a series of frequency distributions. Three different different dependent variables were computed from the judges' recommendations: an initial standard (Standard 1), a second standard (Standard 2) and a "delta" which is the difference obtained by subtracting Standard 1 from Standard 2. Delta represents the change in recommended standards that results from the iterative two-stage process.

In order to examine the association between concomitant variables and recommended standards, Standard 2 and delta were computed for each level of each concomitant variable. Because there were many instances in which the sample size was very small for one or more of the five levels of these variables, it was decided to collapse the "Strongly Agree" and "Agree" categories, the "Strongly Disagree" and "Disagree" categories and to omit the

neutral "Don't Know" categories. The mean recommended standard was computed at each of the two resulting levels, a one-way ANOVA was performed, and the significance of F was computed together with an effect size (the difference between means divided by the within-group standard deviation).

RESULTS AND DISCUSSION

Analyses of Recommended Standards

Standards recommended by each judge were computed by summing their individual item-probability judgements. The resulting recommended standards were on a raw number-of-items correct scale. Separate standards were computed for the initial judgment session (Standard 1) and the second session (Standard 2). A small number of judges were present for either the first or second judgment sessions but not both; if recommended standards from both sessions were not available, data were excluded from further analysis. Table 3 summarizes the recommended standards.

Insert Table 3 about here

The mean, median and standard deviation are reported for each recommended standard; the number of items in each subtest is also provided. Mean initial standards ranged from 11.44 (for the 25-item Mathematics subtest) to 20.38 (for the 45-item Writing subtest). The percentage of items which the judges felt a "minimally competent individual" should answer correctly ranged between 40% (Science subtest) and 50% (the Reading and Social Studies subtests).

As was the case in a pilot study reported earlier (Jaeger & Busch, 1984a), there were no consistent changes in the mean values of recommended standards from the initial to the second judgment session. Three subtest means increased over time, while four decreased. In the pilot study we noted that although there was no change in central tendency when the two standards were compared, there was a decrease in the variability of recommended standards. In the present study, the variances decreased following the provision of information and discussion in each of the subtest panels; however, when hypotheses of equal variances in dependent samples was tested (Glass & Hopkins, 1984, p. 268), significant decreases were found only for the Reading, Writing and Social Studies subtests.

Analyses of Judges' Responses to Questionnaire

As noted above, the final element of the data-collection procedure had participants report their attitudes and their perceptions of the standard-setting activities. A single evaluation questionnaire, consisting of 25 multiple-choice items, was administered to all participants. Responses to evaluation items were recorded on computer-readable scoring sheets which also contained space for additional comments. Response choices for each item were: SD-Strongly Disagree, D--Disagree, DK--Don't Know, A--Agree, SA--Strongly Agree. All participants were asked to respond to every item.

For purposes of analysis, items on the Participant's Evaluation Questionnaire were grouped under four headings that reflect the topics addressed by the questionnaire. The tabular results of these analyses are reported in Tables 4-7. Each table contains

an item number, a concise statement of the question asked, a listing of each response option, from to "Strongly Agree," to "Strongly Disagree" and "missing", the frequency of the participants' responses to each option, the percent of participants who responded to each option, and the cumulative percent of responses to each option. Results are reported collectively for all panelists who reviewed multiple-choice tests, irrespective of the particular subtest reviewed. When the responses for specific subtests differed from the total group, those differences were noted and discussed.

Attitudes: The first series of questions to be analyzed pertained to judges' attitudes toward setting standards, toward screening applicants to teacher education programs and toward the use of standardized tests for that purpose (Table 4). Item 2 asked participants to respond to the statement that they "did not understand the need to set standards." Ninty percent of the participants disagreed with this statement. Item 4 asked about participants' perceptions of the need to screen candidates for teacher education programs. Ninty percent of the participants disagreed with the statement that "it is not important to screen candidates." Item 5 asked whether the participants thought that standardized testing was an appropriate way to screen candidates for teacher education programs. Almost 60% of the participants agreed with the statement, 23% disagreed, and 17% didn't know. Item 6 asked participants to respond to the statement that the NTE tests were not appropriate for screening applicants to teacher education programs. About 57% of the participants

disagreed with the statement, 27% agreed with it, and 16% indicated that they did not know.

Insert Table 4 about here

In summary, a majority of the participants strongly supported a main objective of the North Carolina Teacher Quality Assurance Program, the establishment of standards and the screening of candidates to teacher education programs. Also, a majority of the participants agreed that standardized tests are appropriate for this purpose, and that the NTE tests, in particular, are appropriate devices. There were some notable differences for Items 5 and 6 when the responses were analyzed separately by subtest. A greater number of judges in the Writing (70%) and Listening (69%) panels agreed that standardized testing was an appropriate way to screen candidates for admission to teacher education programs; however, fewer judges on the Reading (46%), Science (45%), and Social Studies (50%) panels agreed. A greater number of Listening (76%) panel judges disagreed with the statement that the NTE was <u>not</u> appropriate for screening candidates for admission to teacher education programs; fewer judges on the Mathematics (36%) and Science (43%) panels disagreed with the statement than did the total group.

The second category of questions on the evaluation questionnaire pertained to participants' perceptions of confidence in their standard-setting judgments. Only one item was in this category; the analysis for this item is reported in Table 5.

Item 13 stated that participants were not confident about their session 1 probability estimates. Disagreement with item 13

Insert Table 5 about here

---

suggests higher confidence in the probability estimates during the first standard setting session. About 59% of the participants disagreed with that statement, while about 35% agreed that they were not confident. In summary, while a majority of the participants (59%) agreed that they were confident about the probabilities they estimated, a large minority (35%) did not feel confident. Judges on the individual panels varied in their responses to this question. Mathematics (68%) and Science (71%) judges disagreed at a greater rate (showed more confidence); Writing and Listening panelists expressed less confidence and reported agreement rates of 50% and 45%, respectively.

Understanding: The third set of questions examined participants' understanding of the standard-setting procedures. Analyses for this series of questions are reported in Table 6.

Item 12 asked participants whether they understood the probability scale used in setting standards. About 94% of the participants agreed that they understood the probability scale.

Insert Table 6 about here

Item 15 stated that the participants did not understand the

percents in parentheses (overall item p values converted to percents), therefore disagreement with this statement suggested understanding of the percentage information. About 97% of the participants disagreed with the statement. Item 16 stated that participants understood the numbers in boxes (p-values for examinees with total scores between 30% and 50% correct, converted to percents) given on the mini-test exercise. Almost 95% agreed that they did understand that information. Finally, Item 17 asked if the graph used in the standard-setting sessions made sense. Over 94% of the participants agreed that the graph used for the standard-setting sessions made sense. In summary, the participants evaluations strongly supported their understanding of the standard-setting procedures. There did not appear to be major differences in understanding of procedures when different panels were compared.

Utilization: The final series of questions to be analyzed pertained to participants' use of and influence by the information provided, and the discussion in which they participated during the standard-setting procedures. Results of these analyses are presented in Table 7.

Item 14 stated that participants utilized information from

Insert Table 7 about here

---

the mini-test exercise; about 81% of the participants agreed that they used this information. Item 20 stated that data on the

percentages of all applicants answering the question correctly had no influence on judges' second session recommendations. About 32% of judges agreed that the information had no influence and about 64% disagreed with the statement. Item 21 stated that data on the percentage of applicants with scores in the 30-50% range influenced their recommendations. Fifty-eight percent of the participants agreed that the information did influence them, and 37% disagreed with the statement. Item 22 proposed that data on failure rates shown in the graph had no influence on judges' second session recommendations. Forty-nine percent of the participants disagreed suggesting that the information influenced them, while 45% agreed that it had no influence. Item 23 focused on the effect of discussion held during the standard-setting session. Sixty percent of judges agreed that discussion did influence their recommendations, and 33% disagreed. In summary, a majority of participants reported that they were influenced in their standard-setting recommendations by these sources of information: 1) the mini-test information (81%); 2) data on the percent of all applicants tested in April, 1983 (64%); 3) data on the percent of applicants in the 30-50% range (58%), the graph data (49%), and discussions (60%).

However there were differences among members of different subtest panels in their responses to Items 21, 22, and 23. Item 21 stated that information on the percentage of all applicants with scores in the 30- to 50- range influenced the estimation of probabilities in the second standard-setting session. Judges who evaluated the Reading test agreed at a greater rate (73%); judges on the Listening panel agreed at a lesser rate (48%). Item 22

indicated that the graph did not influence judges' probability estimates. More Mathematics panelists indicated that the graph did influence judgment by disagreeing with the statement (64%), while Science (39%) and Listening (38%) judges tended to disagee at a lower rate. Item 23 stated that the judges' recommendations during the second-standard setting session were influenced by the discussion of items. Literature/Fine Arts (81%), Science (77%), and Reading (73%) judges agreed to the statement more frequently than did the total group. Writing judges (34%) were less likely to agree with the statement.

In conclusion, analysis of responses to the participants' questionnaire, indicated that the participants generally agreed with the need to validate and set standards for the NTE. In addition, a majority of the participants understood the information provided and the procedures used in setting standards. Similarly, a majority of participants felt confident about the judgments they made, and made use of information intended to help them in making recommendations, when such information was available. There were some differences in the responses of panelists who reviewed different subtests. In general, it appeared that judges on the Writing and Listening panels appeared to be more accepting of the use of standardized tests and the NTE in particular, but were less confident of their judgments. Mathematics, Reading and Science panelists were more likely to use the information provided and were more confident of their judgments.

Associations between Concomitant Variables and Recommended NTE Standards

Recommended standards were contrasted for each level of the

concomitant variables. The sample size at individual levels of some of the concomitant variables was very low; for example, on some of the questionnaire items few or no judges "strongly agreed". For this reason, it was decided to combine the "strongly agree" and the "agree" to form a "strongly/agree" category representing general agreement. For the same reason, "Strongly Disagree" and "Disagee" were also combined. The central tendency and variance of the recommended standards in the general "agree" and "disagree" categories were compared by using the analysis of variance. Since there were sixteen categorical variables for each of the subtests and each analysis was conducted using two dependent variables (Standard 2 and Delta) a large number of contrasts were conducted. Therefore there was a danger that some number of tests would be significant due to chance alone. For this reason it was decided that meaning would be attributed to the results only if some minimal number of contrasts were significant and if the results were consistent with each other and if the results made theoretical sense. In fact, 21 of 112 tests that used standard 2 as a dependent variable were significant and 16 of the 112 tests that used Delta as a dependent variable were significant (based on one-way ANOVAs with alpha = .10). The results of this series of analyses are reported in Table 8.

Most of the concomitant variables did not influence recommended test standards significantly. However, one pattern of

---

Insert Table 8 about here

---

results for Standard 2 does seem worthy of some cautious atten-

tion. When the set of judges' attitude variables is considered alone (Items 02, 04, 05, and 06), eleven of the 28 contrasts were found to be significant. Significant contrasts appear most often for the Writing and Science subtests. Detailed results for the eleven significant contrasts are reported in Table 9, which presents the mean standards for each level of each attitude variable as well as an effect size for each category. Effect sizes given by the difference in means divided by the pooled estimate of the population standard deviation. Some of these results (eg. for items 05 and 06 for the Social Studies subtests) must be interpreted cautiously because of a combination of unequal sample size and heteroscedasticity of variances.

Insert Table 9 about here

Both effect sizes reported for Item 02 were positive; Item 02 stated that the judge did <u>not</u> understand the need to set standards. Therefore one who disagreed with the statement was, in effect, agreeing with the need to set standards. We observed higher recommended standards for those judges who agreed with the need to set standards.

Judges who disagreed with Item 04 indicated that they thought it important to screen candidates for teacher education. Three positive effect sizes are reported in Table 9. This indicates that judges who thought screening to be important recommended higher standards than did those who did not think it important.

Three negative effect sizes were found for Item 05. Judges who agreed with the statement that standardized testing was an appropriate way to screen candidates to teacher education recommended higher standards than did those who disagreed. Item 06 stated that the NTE was not an appropriate way to screen candidates. Three negative effect sizes were found. Those who disagreed (i.e. thought that the NTE was appropriate) recommended higher standards.

In general these results are consistent with what we would predict from theory. It is plausible that judges who do not support the objective of setting standards nor the specific use of the NTE would recommend more lenient standards. From the perspective of cognitive dissonance theory, a judge who expresses doubts about (1) setting standards, about (2) standardized testing, and (3) the use of the NTE, would create dissonance for him/herself if a higher (more demanding) standard were to be set; therefore to resolve the dissonance, a lower standard might be recommended.

Relationships between the variation in recommended standards and the various concomitant variables were also investigated. A total of 15 of 112 contrasts in variances were found significant when the equality of variances was tested with an alpha level of 0.10. There appears to be no systematic relationship between these concomitant variables and variation in judges' recommended test standards.

In general we must conclude that relationships between most characteristics of judges and their recommended test standards are modest. If the relationships between judges' attitudes

and their recommended test standards found in this study are confirmed in future research, certain suggestions seem appropriate. First, our earlier recommendations (Jaeger & Busch, 1984b) that a variety of subpopulations with legitimate interest and expertise in the field in which standards are being set be sampled for establishing standards, seem appropriate. Groups selected on that premise would likely include individuals who vary in their general attitudes toward the screening of candidates, in their attitude regarding the use of standardized tests, etc. We believe that a healthy variation in such attitudes would serve the best interests of those who wish to establish test standards. Second, in the procedures used in the present study, group leaders spoke about the purpose of the North Carolina Teacher Quality Assurance program, the characteristics of the NTE, etc. Those procedures were designed to present information and were not designed to change attitudes. We would recommend that once judges have been selected (with varying attitudes toward testing etc.), they be provided with fundamental information on the purposes of, and rationale for the the use of tests in selection, but that no attempt be made to modify existing judges' attitudes. These recommendations would appear to be consistent with the conclusions of Fitzpatrick (1984).

Finally, although judges' perceptions of their use of information on the test performance of examinees were not consistently related to their recommended test standards, we strongly suggest that such information be provided in standard-setting procedures. The vast majority of judges reported that

such data influenced their judgments. Inconsistencies in directions of influence are a natural consequence of valid differences in judgment and are not a basis for eliminating information that produces better informed judgment. As we have noted in earlier research (Jaeger & Busch, 1984), using representative samples of interested, informed judges is essential to effective, defensible standard setting.

REFERENCES

Asch, S. (1951). Effects of group pressure upon the modification and distortion of judgements. In H. Guetzkow (Ed.), Groups, leadership and men. Pittsburg: Carnegie Press.

Brandstatter, H., Davis, J.H., & Stocker-Kreichgauer, G. (1982). Group Decision Making. New York, New York: Academic Press.

Brown, R. (1965). Social Psychology. Toronto, Ca: Collier - Macmillan.

Busch, J., Jaeger, R., Pratto, D., & Sullivan, J. (1983, June). Phase II of a validation and standard setting study of the General Knowledge and Communications Skills sub-tests of the National Teacher Examinations. Vol. 1: Results of pilot investigation. Report to the North Carolina State Board of Education. Greensboro, North Carolina: Center for Educational Research and Evaluation, UNC-Greensboro.

Fitzpatrick, A. (1984). Social influences in standard setting: The effect of group interaction on individual's judgments. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Glass, G. and Hopkins, K. (1984). Statistical methods in psychology and education. Englewood: Prentice Hall.

Jaeger, R. and Busch, J. (1984a, March). A validation and standard setting study of the General Knowledge and Communications Skills sub-tests of the National Teacher Examinations: Final Report. Report to the North Carolina State Board of Education. Greensboro, North Carolina: Center for Educational Research and Evaluation, UNC-Greensboro.

Jaeger R. and Busch, J. (1984b, April). The effects of a delphi modification of the Angoff-Jaeger standard-setting procedure on standards recommended for the National Teacher Examinations. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Myers, D. & Lamb, H. (1976). The group polarization phenomenon. Psychological Bulletin, 83, 602-627.

Popham, J.W. and Yalow, E.S. (1984, April). Standard Setting Options for Teacher Competency Tests. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Sherif, M. (1947). Group influences upon the formation of norms and attitudes. In Newcomb, T. & Hartley, E. L. (Eds.) Readings in Social Psychology (1st Ed.). New York: Holt.

Stephan, C. & Stephan, W. (1985). Two social psychologies. Homewood, Illinois: Dorsey Press.

Stoner, J. (1961). A comparison of individual and group decisions including risk. Unpublished masters thesis. School of Industrial Management, M.I.T.

Swap Associates. (1984). Group decision making. Beverly Hills, CA.:Sage Publications.

Table 1

Number of Judges on Each Panel Who were Public School Teachers or College/University Faculty

Type of Judge	NTE Communication Skills Subtest		
	Listening	Writing(MC)	Reading
Public School	17	14	5
College/University	25	36	21
Total	42	50	26

Type of Judge	NTE General Knowledge Subtests			
	Science	Mathematics	Social Studies	Literature Fine Arts
Public School	8	8	8	14
College/University	23	17	22	23
Total	31	25	30	37

Table 2

Concomitant and Dependent Variables

Judges' Characteristics

Gender
Position (Public School Teacher, University/College Professor)
Subtest Panel (1=Reading, 2=Writing, 3=Listening, 4=Mathematics, 5=Science, 6=Social Studies, 7=Literature/Fine Arts)

Judges' Attitude Toward Screening and Standard Setting

2. I do not understand the need to set standards on the NTE tests for use with teacher education applicants. (R)
4. I do not think it is important to screen candidates for admission to North Carolina teacher education programs. (R)
5. I think that standardized testing is an appropriate way to screen candidates for admission to North Carolina teacher education programs.
6. I do not think that the NTE tests are appropriate for screening candidates for admission to North Carolina teacher education programs. (R)

Judges' Confidence in Recommended Standards

13. I do not feel confident about my first (before lunch) estimates of probabilities of correct answers for minimally competent applicants to teacher education programs. (R)

Judges Understanding of Information and Procedures Utilized

12. I understand the meaning of the ten-point probability scale used in the standard setting procedure.
15. I did not understand the meaning of the percents in parenthesis beside the items on the Mini-test exercise; [p-values for the entire applicant population. (R)
16. I fully understood the meaning of the numbers in boxes beside the items on the Mini-test exercise. [p-values for those with total scores in the 30% to 50% range]
17. The graph distributed at the beginning of the second standard-setting session made sense to me.

Judges' Utilization of Information Provided

14. The Mini-test exercise provided information that I used in formulating my second set of estimated probabilities for minimally competent applicants.
20. Data on the percentage of all applicants who answered each test item correctly did not influence my estimated probabilities during the second standard setting session. (R)
21. Data on the percentage of applicants with scores in the 30- to 50-percent range who answered each test item correctly, influenced my estimated probabilities during the second standard-setting session.

Table 2 (continued)

22. Data in the graph I received at the beginning of the second standard-setting session, on the percentage of North Carolina applicants who would fail the test, did not influence my estimates during that session. (R)

23. My recommendations during the second standard-setting session were influenced by the discussion of items during that session.

Criterion Variables

First standard recommended using the Angoff method with no background information provided (Standard 1).

Second standard recommended using the Angoff method following provision of information and discussion (Standard 2).

Difference between first and second standard, (Standard 2-Standard 1 = Delta).

(R) denotes a negatively stated item for which the scale was reversed.

TABLE 3 Mean, Median, and Standard Deviation for Standard 1, Standard 2, Change in Standard and Test Length for each NTE subtest

Subtest and Test Length		Standard 1			Standard 2			Change		
		X	Mdn	s	X	Mdn	s	X	Mdn	s
1. Reading	(30)	15.07	14.11	5.16	14.19	14.35	3.47	-.88	-.30	2.98
2. Writing	(45)	20.38	20.20	5.35	20.72	20.58	3.51	.34	-.15	4.90
3. Listening	(40)	19.46	19.90	5.00	19.32	19.60	4.17	-.14	.00	4.72
4. Mathematics	(25)	11.44	11.78	2.36	10.33	9.98	2.05	-1.11	-.70	2.07
5. Science	(30)	11.89	12.65	2.51	12.56	12.85	2.32	.67	.50	1.69
6. Social Studies	(30)	14.58	14.55	4.11	14.40	14.15	2.52	-.18	.11	3.43
7. Literature/ Fine Arts	(35)	15.63	15.03	4.73	16.36	15.78	4.05	.73	.09	3.09

Table 4

Responses of Judges to Items Assessing Attitude Toward Screening and Standard Setting

ITEM02 I DID NOT UNDERSTAND THE NEED TO SET NTE STANDARDS

VALUE LABEL	VALUE	FREQUENCY	PERCENT	VALID PERCENT	CUM PERCENT
STRONGLY DISAGREE	1	145	59.4	60.9	60.9
DISAGREE	2	68	27.9	28.6	89.5
DONT KNOW	3	3	1.2	1.3	90.8
AGREE	4	19	7.8	8.0	98.7
STRONGLY AGREE	5	3	1.2	1.3	100.0
	.	4	1.6	MISSING	
	0	2	.8	MISSING	
	TOTAL	244	100.0	100.0	

ITEM04 IT IS NOT IMPORTANT TO SCREEN CANDIDATES FOR ADMISSION

VALUE LABEL	VALUE	FREQUENCY	PERCENT	VALID PERCENT	CUM PERCENT
STRONGLY DISAGREE	1	159	65.2	66.3	66.3
DISAGREE	2	57	23.4	23.8	90.0
DONT KNOW	3	9	3.7	3.8	93.8
AGREE	4	7	2.9	2.9	96.7
STRONGLY AGREE	5	8	3.3	3.3	100.0
	.	4	1.6	MISSING	
	TOTAL	244	100.0	100.0	

ITEM05 STANDARDIZED TESTING IS AN APPROPRIATE WAY TO SCREEN CANDIDATES

VALUE LABEL	VALUE	FREQUENCY	PERCENT	VALID PERCENT	CUM PERCENT
STRONGLY DISAGREE	1	19	7.8	7.9	7.9
DISAGREE	2	36	14.8	15.0	22.9
DONT KNOW	3	42	17.2	17.5	40.4
AGREE	4	114	46.7	47.5	87.9
STRONGLY AGREE	5	29	11.9	12.1	100.0
	.	4	1.6	MISSING	
	TOTAL	244	100.0	100.0	

Table 4 continued

ITEM06 NTE ARE NOT APPROPRIATE DEVICES FOR SCREENING TEACH. CANDIDATES

VALUE LABEL	VALUE	FREQUENCY	PERCENT	VALID PERCENT	CUM PERCENT
STRONGLY DISAGREE	1	37	15.2	15.5	15.5
DISAGREE	2	98	40.2	41.0	56.5
DONT KNOW	3	39	16.0	16.3	72.8
AGREE	4	40	16.4	16.7	89.5
STRONGLY AGREE	5	25	10.2	10.5	100.0
	.	4	1.6	MISSING	
	0	1	.4	MISSING	
	TOTAL	244	100.0	100.0	

Table 5

Responses of Judges to Items Assessing Confidence in Recommendations

ITEM13 I AM NOT CONFIDENT ABOUT MY FIRST ESTIMATES OF PROBABILITIES

VALUE LABEL	VALUE	FREQUENCY	PERCENT	VALID PERCENT	CUM PERCENT
STRONGLY DISAGREE	1	29	11.9	12.1	12.1
DISAGREE	2	112	45.9	46.7	58.7
DONT KNOW	3	14	5.7	5.8	64.6
AGREE	4	73	29.9	30.4	95.0
STRONGLY AGREE	5	12	4.9	5.0	100.0
	.	4	1.6	MISSING	
	TOTAL	244	100.0	100.0	

Table 6

Responses of Judges to Items Assessing Understanding of Information

ITEM12 I UNDERSTOOD THE 10 PT PROBABILITY SCALE

VALUE LABEL	VALUE	FREQUENCY	PERCENT	VALID PERCENT	CUM PERCENT
STRONGLY DISAGREE	1	1	.4	.4	.4
DISAGREE	2	5	2.0	2.1	2.5
DONT KNOW	3	7	2.9	2.9	5.4
AGREE	4	143	58.6	59.8	65.3
STRONGLY AGREE	5	83	34.0	34.7	100.0
	.	4	1.6	MISSING	
	0	1	.4	MISSING	
	TOTAL	244	100.0	100.0	

ITEM15 I DID NOT UNDERSTAND PERCENTS IN PAREN. ON THE MINI-TEST

VALUE LABEL	VALUE	FREQUENCY	PERCENT	VALID PERCENT	CUM PERCENT
STRONGLY DISAGREE	1	119	48.8	49.6	49.6
DISAGREE	2	113	46.3	47.1	96.7
DONT KNOW	3	1	.4	.4	97.1
AGREE	4	5	2.0	2.1	99.2
STRONGLY AGREE	5	2	.8	.8	100.0
	.	4	1.6	MISSING	
	TOTAL	244	100.0	100.0	

ITEM16 I FULLY UNDERSTOOD THE PERCENTS IN BOXES ON MINI-TEST

VALUE LABEL	VALUE	FREQUENCY	PERCENT	VALID PERCENT	CUM PERCENT
STRONGLY DISAGREE	1	2	.8	.8	.8
DISAGREE	2	9	3.7	3.8	4.6
DONT KNOW	3	2	.8	.8	5.4
AGREE	4	115	47.1	47.9	53.3
STRONGLY AGREE	5	112	45.9	46.7	100.0
	.	4	1.6	MISSING	
	TOTAL	244	100.0	100.0	

ITEM17 THE GRAPH MADE SENSE TO ME

VALUE LABEL	VALUE	FREQUENCY	PERCENT	VALID PERCENT	CUM PERCENT
STRONGLY DISAGREE	1	6	2.5	2.5	2.5
DISAGREE	2	7	2.9	2.9	5.4
AGREE	4	124	50.8	51.7	57.1
STRONGLY AGREE	5	103	42.2	42.9	100.0
	.	4	1.6	MISSING	
		244	100.0	100.0	

Table 7

Responses of Judges to Items Assessing Judges' Use of Information

ITEM14 MINI-TEST PROVIDED INFOR. I USED IN FORMULATING SECOND PROBABIL.

VALUE LABEL	VALUE	FREQUENCY	PERCENT	VALID PERCENT	CUM PERCENT
STRONGLY DISAGREE	1	9	3.7	3.8	3.8
DISAGREE	2	25	10.2	10.5	14.2
DONT KNOW	3	12	4.9	5.0	19.2
AGREE	4	161	66.0	67.4	86.6
STRONGLY AGREE	5	32	13.1	13.4	100.0
	.	4	1.6	MISSING	
	0	1	.4	MISSING	
	TOTAL	244	100.0	100.0	

ITEM20 THE % ALL APPLICANTS ANSWERING CORRECTLY NOT INFLUENCE ME

VALUE LABEL	VALUE	FREQUENCY	PERCENT	VALID PERCENT	CUM PERCENT
STRONGLY DISAGREE	1	18	7.4	7.5	7.5
DISAGREE	2	136	55.7	56.7	64.2
DONT KNOW	3	10	4.1	4.2	68.3
AGREE	4	65	26.6	27.1	95.4
STRONGLY AGREE	5	11	4.5	4.6	100.0
	.	4	1.6	MISSING	
	TOTAL	244	100.0	100.0	

ITEM21 % APPLICANTS IN 30-50 % RANGE ANSWER CORRECTLY INFLUENCED ME

VALUE LABEL	VALUE	FREQUENCY	PERCENT	VALID PERCENT	CUM PERCENT
STRONGLY DISAGREE	1	19	7.8	7.9	7.9
DISAGREE	2	69	28.3	28.9	36.8
DONT KNOW	3	12	4.9	5.0	41.8
AGREE	4	126	51.6	52.7	94.6
STRONGLY AGREE	5	13	5.3	5.4	100.0
	.	4	1.6	MISSING	
	0	1	.4	MISSING	
	TOTAL	244	100.0	100.0	

Table 7 continued

ITEM22 GRAPH DID NOT INFLUENCE MY SECOND JUDGMENTS

VALUE LABEL	VALUE	FREQUENCY	PERCENT	VALID PERCENT	CUM PERCENT
STRONGLY DISAGREE	1	20	8.2	8.3	8.3
DISAGREE	2	98	40.2	40.8	49.2
DONT KNOW	3	15	6.1	6.3	55.4
AGREE	4	91	37.3	37.9	93.3
STRONGLY AGREE	5	16	6.6	6.7	100.0
	.	4	1.6	MISSING	
	TOTAL	244	100.0	100.0	

ITEM23 MY SECOND JUDGMENTS WERE INFLUENCED BY THE DISCUSSION

VALUE LABEL	VALUE	FREQUENCY	PERCENT	VALID PERCENT	CUM PERCENT
STRONGLY DISAGREE	1	19	7.8	7.9	7.9
DISAGREE	2	60	24.6	25.1	33.1
DONT KNOW	3	16	6.6	6.7	39.7
AGREE	4	131	53.7	54.8	94.6
STRONGLY AGREE	5	13	5.3	5.4	100.0
	.	4	1.6	MISSING	
	0	1	.4	MISSING	
	TOTAL	244	100.0	100.0	

Table 8

Summary of Breakdowns of Standard 2 and Delta by Concomitant Variable Characteristics; Significant F Values.

Concomitant Variab.	Standard 2 Analysis for Subtest							Delta Analysis for Subtest						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Background														
Gender					X					X				X
Position			X						X		X			X
Attitude														
Item 02		X			X								X	X
Item 04	X	X			X			X					X	
Item 05		X			X	X								
Item 06		X			X	X								
Confidence														
Item 13											X			X
Understanding														
Item 12									X					
Item 15			X											
Item 16														X
Item 17														
Utilization														
Item 14										X				
Item 20	X													
Item 21			X											
Item 22		X				X						X		
Item 23		X				X					X			

X = Denotes significant contrast ($p<.10$)

Table 9

Relationship Between Standard 2 and Judges Attitude: Means, and Effect Sizes

Summary for Significant Contrasts	Judge Attitude Variables Item 02	Item 04	Item 05	Item 06
X - Strongly/Disagree	21.19	21.08	19.14	21.47
X - Strongly/Agree	18.22	16.63	21.49	19.08
Effect Size	+.87	+1.32	-.67	+.70
Subtest	Writing	Writing	Writing	Writing
X - Strongly/Disagree	12.85	12.90	11.54	13.23
X - Strongly/Agree	8.69	8.69	13.32	11.36
Effect Size	+1.98	+2.12	-.79	+.81
Subtest	Science	Science	Science	Science
X - Strongly/Disagree			13.37	15.49
X - Strongly/Agree			15.57	12.59
Effect Size			-.92	+1.28
Subtest			Social St.	Social St.
X - Strongly/Disagree				
X - Strongly/Agree				
Effect Size				
Subtest				
X - Strongly/Disagree		14.75		
X - Strongly/Agree		11.09		
Effect Size		+1.12		
Subtest		Reading		